MESRGAN ON SIGN LANGUAGE IMAGE RECOGNITION

CHAN JYH HUAH (A0178199H)

HUANG FUXING (A0163461J)

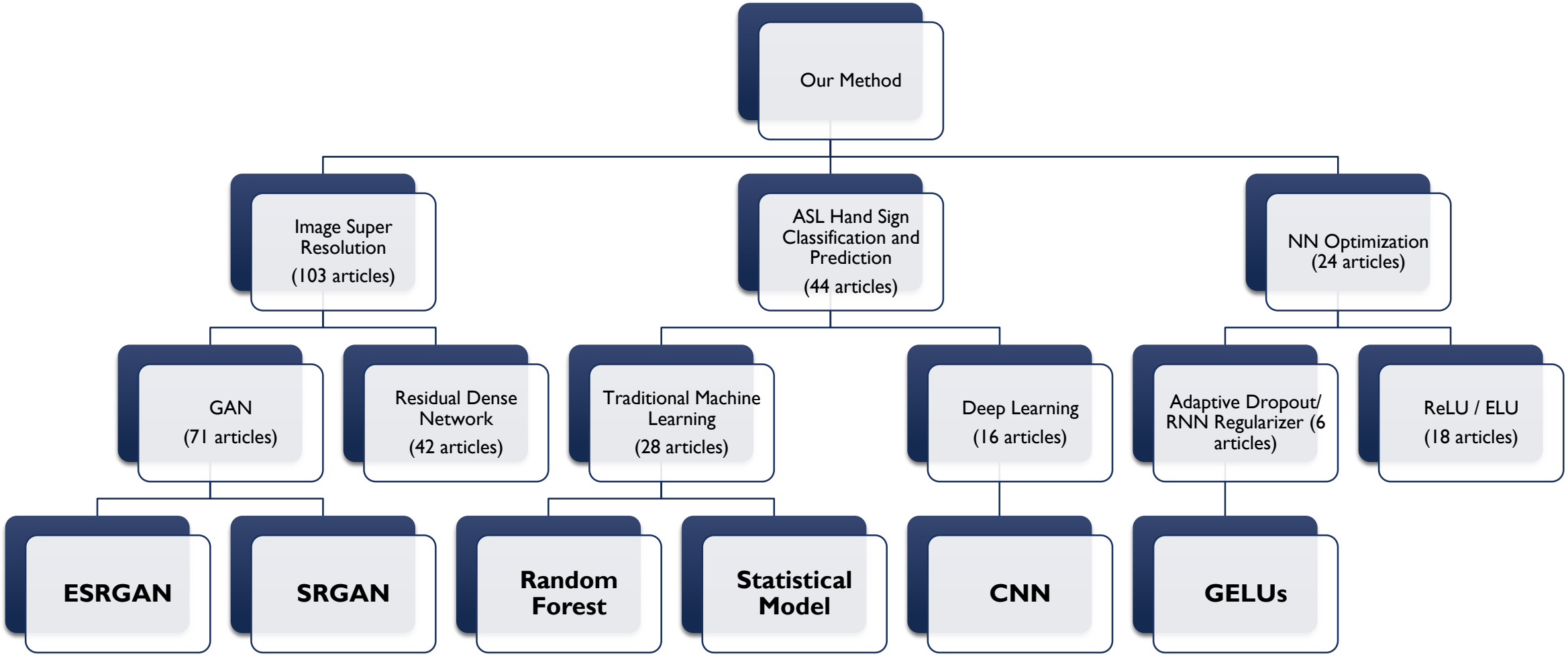
TAN CHEE WEI (A0179723U)

WOO CHEE YOONG (A0178344X)

OUR DATASET

- The dataset format is patterned from the classic MNIST dataset, which is a popular benchmark for image-based machine learning
- The training data (27,455 cases) and test data (7,172 cases) will be applied to K-Fold validation using train-test split of 70:30
- Have 784 columns, each column represent a pixel value, one row data represent a single 28x28 pixel image, with grayscale values between 0-255
- one-to-one map for each alphabetic letter A-Z, so in actual dataset, the labels are 0-25, excluding 9-J & 25-Z.

LITERATURE REVIEW



APPROACH 1: USING CNN

CNN Model with five layers

Two convolution layers

Two max pooling layers

One fully connected layer

- Parameter tuning

Learning rate: 0.4, 0.6, 0.8, 1.0, 1.2

Number of iterations: 20, 40, 60, 80, 100

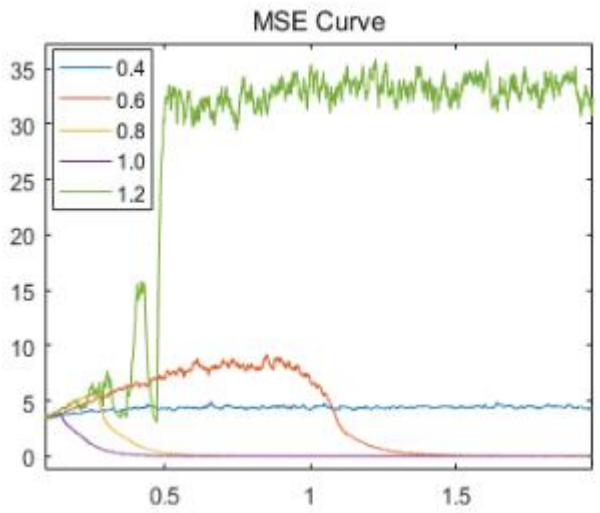
Batch size: 20, 50, 100, 150, 300

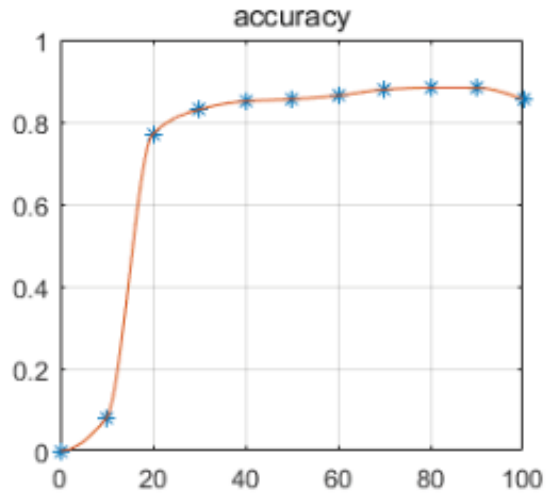
Zhao, Y., & Wang, L. (2018, November). The Application of Convolution Neural Networks in Sign Language Recognition. In 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP) (pp. 269-272). IEEE.

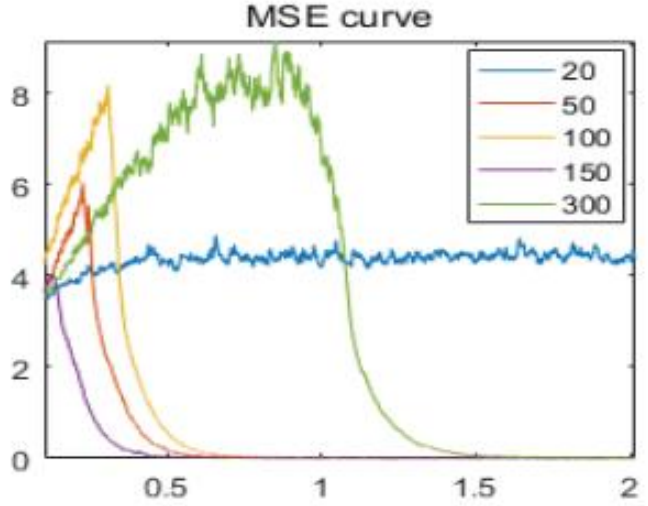
APPROACH 1: USING CNN

Optimized Hyperparameters CNN

- Learning rate of 0.8, Number of iteration of 80 and batch size of 150 was selected







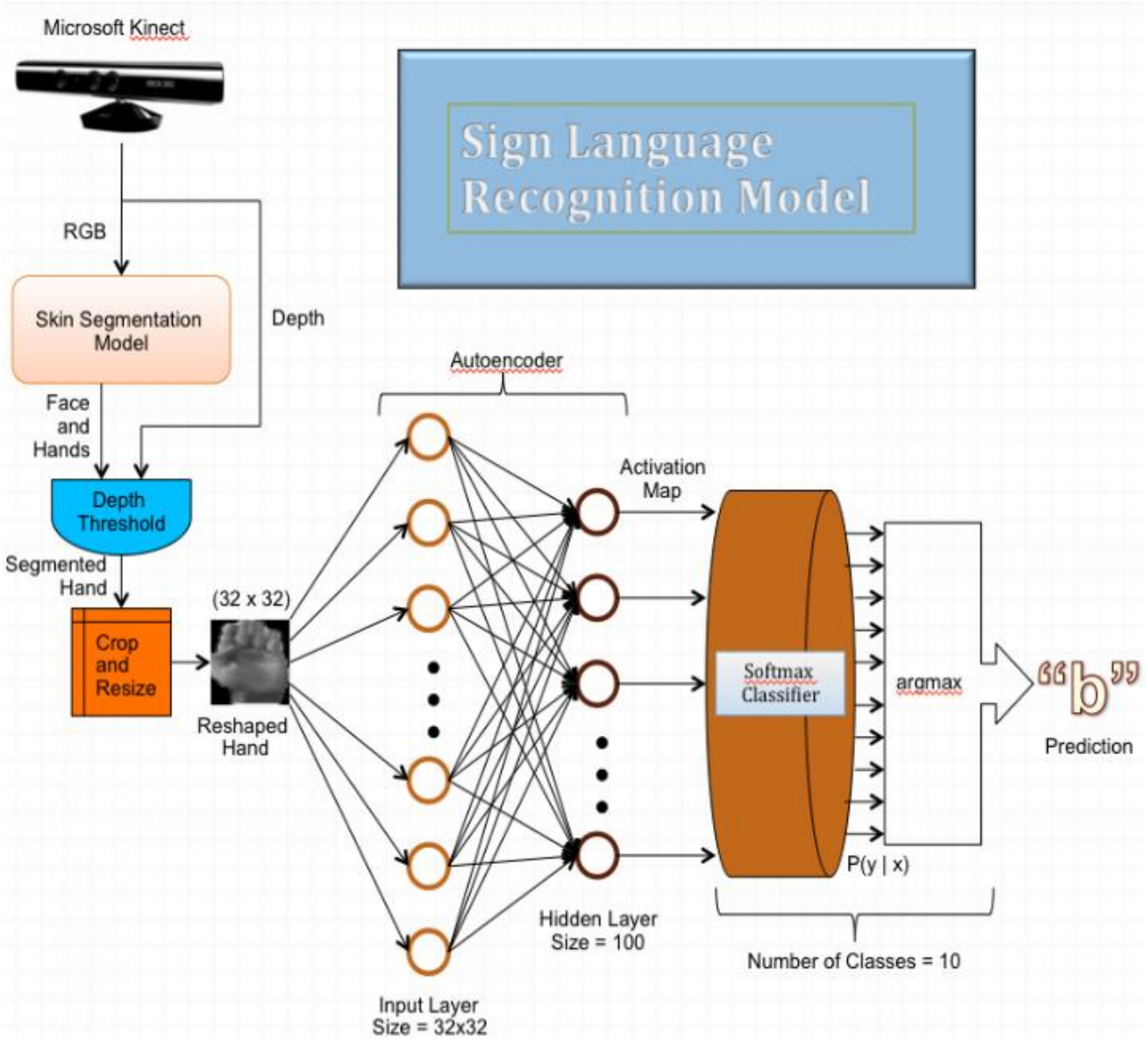
APPROACH II: UNSUPERVISED FEATURE LEARNING + CNN

Segmentation Methods

- Edge segmentation
- Skin Color segmentation

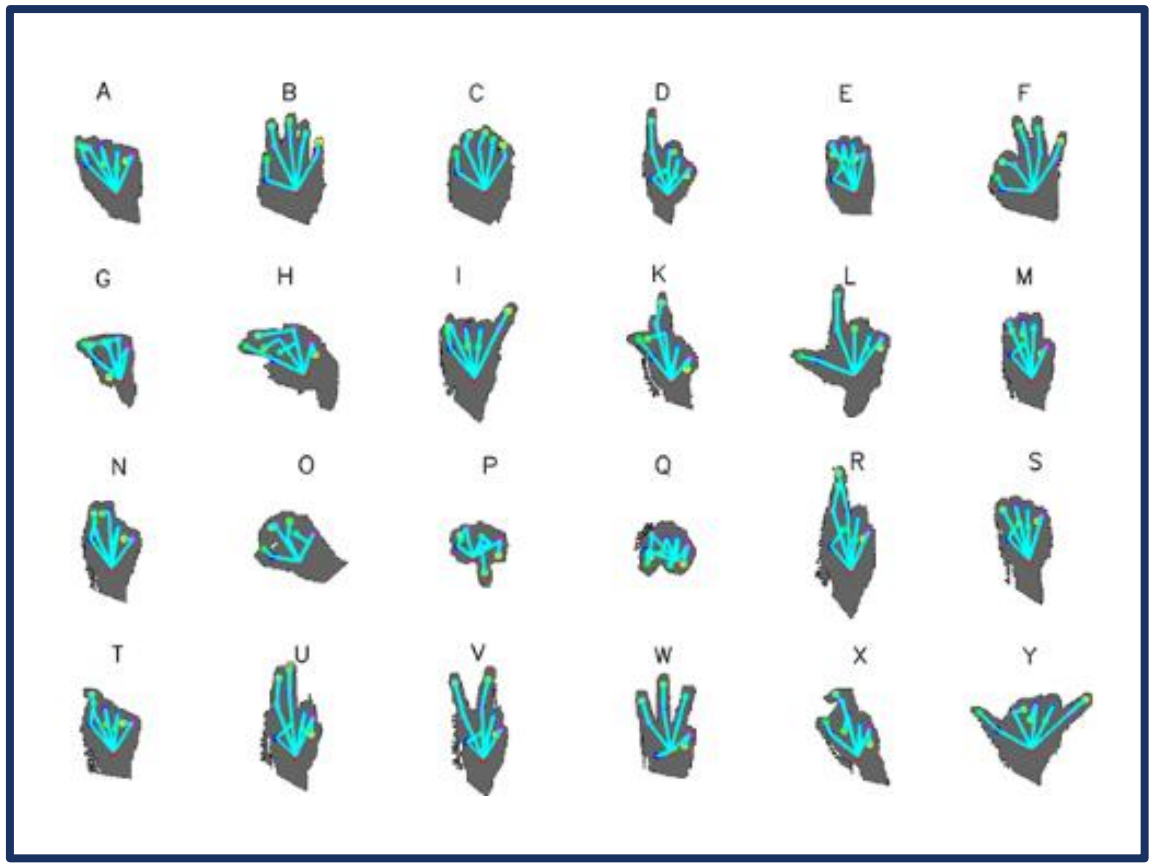
Feature Learning and classification

- Sparse autoencoder
- L-BFGS optimize cost function



Chen, Justin. *Sign Language Gesture Recognition with Unsupervised Feature Learning* , 2011, cs229.stanford.edu/proj2011/ChenSenguptaSundaram-SignLanguageGestureRecognitionWithUnsupervisedFeatureLearning.pdf.

APPROACH III: RANDOM FOREST

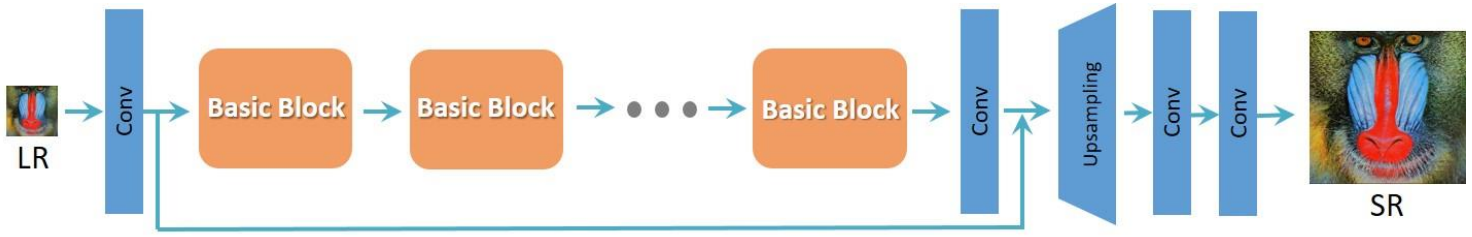


- Feature Extraction
  - Evenly distribute scheme
  - Distance Adaptive Scheme
  - Per pixel classifier
- Gesture Recognition
  - Joint localization
  - Kinematic constraints

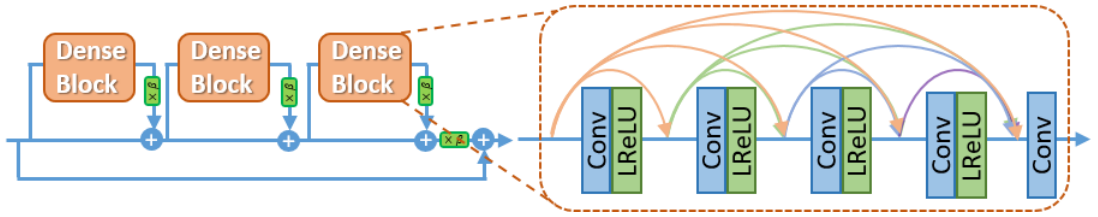
Dong, C., Leu, M. C., & Yin, Z. (2015). American sign language alphabet recognition using microsoft kinect. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44-52).

OUR APPROACH: ESRGAN (IMPROVED OVER THE SRGAN MODEL)

- A novel architecture containing several RRDB blocks without Batch Normalization layers.
- Adopt a deeper model using Residual-in-Residual Dense Block (RRDB) without batch normalization layers.
- Employ Relativistic average GAN instead of the vanilla GAN, which learns to judge whether one image is more realistic than another, guiding the generator to recover more detailed textures.
- Enhanced the perceptual loss by using the features before activation, which offer stronger supervision and thus restore more accurate brightness and realistic textures



Residual in Residual Dense Block (RRDB)



Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 0-0).

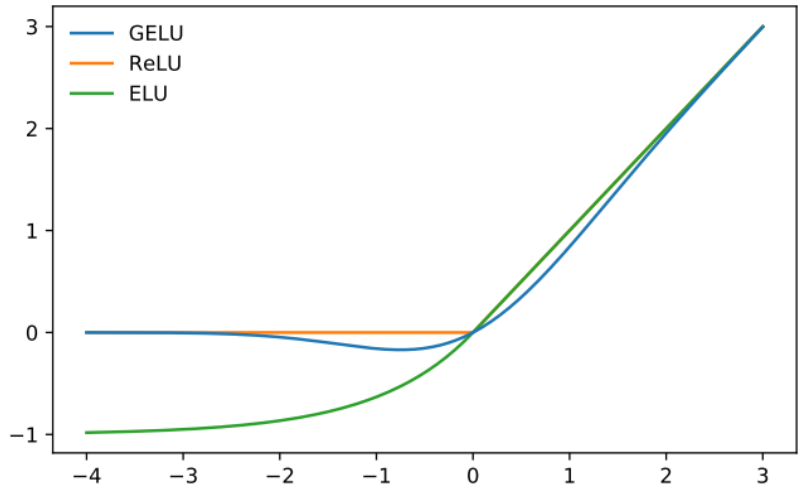
OUR APPROACH: GAUSSIAN ERROR LINEAR UNIT (GELU)

- A high-performing neural network activation function
- Randomly applies zero map to a neuron's input
- Weight inputs by their magnitude rather than by their sign as in RELU's
- Performance improvement across computer vision, natural language processing and speech tasks

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x).$$

Which we can approximate with

$$0.5x(1 + \tanh[\sqrt{2/\pi}(x + 0.044715x^3)])$$



Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

OUR APPROACH

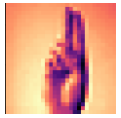
IMAGE OPTIMIZATION USING ESRGAN

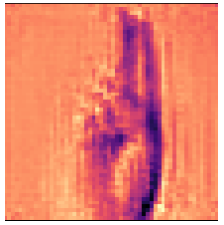
Original Image dataset (28*28 pixel)

ESRGAN Image optimization (i.e. rescale factor = 2)

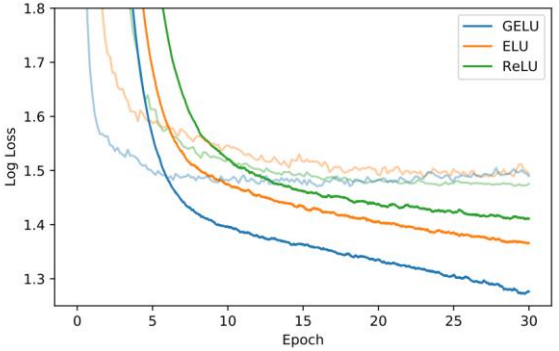
CNN Optimized Model (Activation Function = GELU)

Improved accuracy in ASL Sign language recognition





$$\text{GELU}(x) = xP(X \leq x) = x\Phi(x).$$



COMPARISON OF METHODS

	APPROACH	ACCURACY
1	Convolutional Neural Network	89.32%
2	Unsupervised Feature Learning + CNN	98.0%
3	Random Forest	90%
4	Baseline CNN	85.72%
5	ESRGAN + CNN	92.47%
6	Our Method (MESRGAN)	93.63%

FUTURE WORKS

- Increase the scaling factor of ESRGAN from x2 to x4 or x8, which should increase the image resolution from 56x56 to become 112x112 or 224x224. Theoretically provide more features for machine learning model train/test
- Replace the existing CNN model to VGG16 or ResNet50 or DCNN, which can further improve the learning accuracy
- Trained using video of sign gestures instead of static images which can bring the work closer to real time application of ASL interpretation.

Q&A